12-2018

Root-n consistency of intercept estimators in a binary response model under tail restrictions

Lili TAN
Yunnan University

Yichong ZHANG
Singapore Management University, yczhang@smu.edu.sg

Root-n Consistency of Intercept Estimators in a Binary Response Model Under Tail Restrictions*

Lili Tan† Yichong Zhang‡

March 7, 2017

Abstract

The intercept of the binary response model is irregularly identified when the supports of both the special regressor V and the error term ε are the whole real line. This leads to the estimator of the intercept having potentially a slower than $\sqrt{n}$ convergence rate, which can result in a large estimation error in practice. This paper imposes addition tail restrictions which guarantee the regular identification of the intercept and thus the $\sqrt{n}$-consistency of its estimator. We then propose an estimator that achieves the $\sqrt{n}$ rate. Finally, we extend our tail restrictions to a full-blown model with endogenous regressors.

Keywords: Extremal quantile, Tail index

JEL codes: C13, C14, C25

*We thank Shakeeb Khan, Federico Bugni, Arnaud Maurel, Xavier D'Haultfœuille, the editor, three anonymous referees, and the participants in the 2014 Econometric Society China Summer Meeting and the Duke microeconometrics lunch group for their comments. All remaining errors are ours.

†Yunnan University. E-mail address: tanlili@ynu.edu.cn

‡Corresponding author. Singapore Management University. E-mail address: yczhang@smu.edu.sg.

1 Introduction

In the seminal works of Lewbel (1997) and Lewbel (2000), the intercept of the binary response model is identified and estimated with the aid of a special regressor V. When V has compact support, Lewbel (1997), Lewbel (2000), and Lewbel and Schennach (2007) have shown that the estimator is $\sqrt{n}$-consistent. However, for the sake of identification, the compactness of the support of V implies that either ε also has a compact support (Lewbel, 1997, 2000) or ε has tail symmetry (Magnac and Maurin, 2007). The former condition excludes the basic logit and probit models, as pointed out by Lewbel (1997) and Lewbel (2000). The latter condition depends upon the unknown intercept value being identified and thus is not generic. When V has unbounded support, the $\sqrt{n}$-consistency has been established based on high-level assumptions on the bias and variance of the estimator. See, for example, Lewbel (1997), Lewbel (2000), and Stoker (1991). These high level assumptions do not hold in general because Khan and Tamer (2010) showed the intercept is irregularly identified and cannot be $\sqrt{n}$-consistently estimated without additional tail restrictions. In addition, Khan and Tamer (2010) pointed out that the relative thickness of the tails of V and ε plays the key role of determining the convergence rate, but they did not provide sufficient conditions for the $\sqrt{n}$-consistency.

This paper provides additional tail restrictions that are sufficient for the regular identification of the intercept. We then propose a feasible estimator of the intercept and show it is $\sqrt{n}$-consistent under the restrictions. We also provide another set of restrictions on the tails which ensures the nonexistence of any $\sqrt{n}$-consistent estimator of the intercept.

The tail restrictions we impose for the regular identification basically require that the tail of the special regressor V is thicker than the tail of the error term ε, which is in line with Khan and Tamer (2010). In one particular case, we show the intercept is $\sqrt{n}$-consistently estimable if the unobservable ε has rapidly varying tails (e.g., normal distribution) and the special regressor V has regularly varying tails (e.g., T distributions with any degree of freedom). This result extends the previous results of Khan and Tamer (2010) that if V has infinite variance, there exists a $\sqrt{n}$-consistent estimator of the intercept.

We build our estimator by trimming based on extremal quantiles of the special regressor. Trimming has been used to estimate the binary response model, GMM with heavy-tailed data, and average treatment effect by Yang (2015), Hill and Renault (2010), and Chaudhuri and Hill (2015), respectively. Hill and Renault (2010) and Chaudhuri and Hill (2015) proposed to trim the whole estimand, which is not feasible in our case because our estimand contains the density of the special regressor, which is unknown and should be estimated. Instead, we trim the estimand based on the extremal quantile of the special regressor. In addition, in contrast to the trimming used in Yang (2015), our trimming scheme is driven by data.

The rest of the paper is organized as follows. Section 2 defines a simple model. Section 3 investigates the $\sqrt{n}$-consistency of an estimator of the intercept in the simple model. Section 4 extends the simple

model to a general one, considered in Dong and Lewbel (2015), which incorporates endogenous regressors X and instrumental variables Z. Section 5 shows a brief simulation, and Section 6 contains the conclusion. All proofs are stated in the Appendix and an online supplement.

2 The Model

We consider the binary response model

$$Y_i = \mathbb{1}\{V + X_i'\beta - \varepsilon_i \geq 0\},$$

where the coefficient in front of V is normalized to 1 and covariates X include the constant. Identification and estimation of semiparametric binary choice models have been widely investigated in the literature. If covariates X are independent of ε or satisfy an index condition, the coefficients of the continuously distributed regressors can be estimated by the method of average derivatives proposed by Powell, Stock, and Stoker (1989) and Stoker (1991). Han (1987), Ichimura (1993) and Klein and Spady (1993) further estimated coefficients of the discrete covariates but not the intercept. Furthermore, heteroskedasticity, which is important for economic applications, is ruled out by the full independence assumption. Manski (1975), Manski (1985), and Horowitz (1992) considered the conditional quantile independence, which allows for heteroskedasticity. However, the maximum-score-type estimators proposed by the above three papers are not $\sqrt{n}$-consistent, although the convergence rate of Horowitz's (1992) smoothed maximum score estimator can be made arbitrarily close to $\sqrt{n}$ rate, given sufficient smoothness. In addition, Manski (1988) pointed out a conditional mean restriction does not identify parameter β. Lewbel (1997) and Lewbel (2000) complemented the conditional mean restriction by assuming V is a "special regressor," such that it is independent of ε given X. Then under certain support conditions, Lewbel (1997) and Lewbel (2000) established the identification of β based on conditional mean restriction, allowing for heteroskedasticity on all covariates X except the special regressor. For identification, we will follow the strategy of Lewbel (1997) and Lewbel (2000).

To simplify the discussion, in this and the next section, we follow Khan and Tamer (2010) and consider the case in which X contains only a constant. Then, the simple model can be written as

$$Y_i = \mathbb{1}\{\alpha + V_i - \varepsilon_i \geq 0\}. \tag{2.1}$$

Section 4 will return to the general model in which X includes both the intercept and other covariates that may be endogenous. Following Lewbel (1997), we make the next assumption.

Assumption 1. *(1)* $\{(\varepsilon_i, V_i)\}_{i=1}^n$ *is i.i.d. (2)* ε_i *and* V_i *have full support* $\Re$. *(3)* $V_i \perp \varepsilon_i$. *(4)* $\mathbb{E}\varepsilon_i = 0$.

V is referred to as the special regressor by Lewbel (1997) and Lewbel (2000) because it is independent

of the unobservable ε and its support is the real line. Assumption 1(4) is the common location normalization.

Under Assumption 1, Lewbel (1997) and Lewbel (2000) showed α is identified as

$$\alpha = \mathbb{E}\left[\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}\right], \tag{2.2}$$

in which $f(\cdot)$ denotes the true density of V.

3 The Semiparametric Estimation

3.1 The Estimator

Since the support of V is the real line, its density vanishes at two tails. Consequently, we face the "zero-denominator" problem. To deal with this problem, we propose an estimator that is the sample analogue of the RHS of (2.2) with a trimming function $\hat{I}_{n,i}$:

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n}\Gamma_{n,i} \quad \text{and} \quad \Gamma_{n,i} = \left[\frac{Y_i - \mathbb{1}\{V_i > 0\}}{\hat{f}(V_i)}\right]\hat{I}_{n,i}, \tag{3.1}$$

in which $\hat{f}(\cdot)$ is a kernel estimator of $f(\cdot)$, $\hat{I}_{n,i} = \mathbb{1}\{V_i \in \hat{S}_n\}$, and $\hat{S}_n = (\hat{l}_n, \hat{r}_n)$. $\hat{I}_{n,i}$ is the feasible trimming function whose infeasible counterpart is $I_{n,i} = \mathbb{1}\{V_i \in S_n\}$, in which $S_n = (l_n, r_n)$ where l_n and r_n are the non-random trimming points dependent upon the sample size n. $\hat{l}_n$ and $\hat{r}_n$ are the estimators of l_n and r_n, respectively. l_n, r_n, $\hat{l}_n$ and $\hat{r}_n$ will be defined later.

3.2 Asymptotic Properties

Assumption 2. *Recall that $f(\cdot)$ is the density of V.*
(1) There exists some constant $A > 0$ such that $f(v)$ is monotonic when $|v| > A$ and is bounded away from 0 on any compact subset of $\Re$.
(2) $f(\cdot)$ is ν-th order continuously differentiable with $\nu \geq 2$. All of its ν-th order derivatives are bounded.
(3) For some nonnegative constant σ, there exist positive constants c_1, c_2, and c_3 such that

$$f(v \pm c_2) \leq c_1 f(v)^{1-\sigma}$$

when $|v| \geq c_3$.

Several comments on these assumptions are in order. First, Assumption 2(2) is common in non-parametric kernel estimation with higher order kernels. Second, both Assumption 2(1) and 2(3) are satisfied by many well-known distributions such as normal distribution, t distribution, and Laplace

distribution. Last, Assumption 2(3) holds when $f(v)$ decays polynomially as $|v| \to \infty$ and the density of V is bounded by some constant c. To see this, note that, by Definition A in Resnick (1987, Section 0.4.1), if $f(v)$ decays polynomially[1] as $|v| \to \infty$, then $\frac{f(v \pm c_2)}{f(v)} \to 1$ for any $c_2 \in \Re$. Therefore, for any $\delta > 0$, there exists c_3 such that

$$\sup_{|v|>c_3} \frac{f(v \pm c_2)}{f(v)} \leq 1 + \delta.$$

Then, for any $\sigma \in [0, 1]$,

$$f(v \pm c_2) \leq \sup_v f(v)^\sigma f(v \pm c_2)^{1-\sigma} \leq c f(v \pm c_2)^{1-\sigma} \leq c(1+\delta) f(v)^{1-\sigma}.$$

Assumption 2(3) holds with $c_1 = c(1+\delta)$, any $c_2 \in \Re$, and some c_3 dependent upon δ and c_2.

When $f(\cdot)$ decays exponentially, we further consider the case that $\log(f(\cdot))$ decays polynomially, e.g., the normal and Laplace density. In this case, $\frac{\log(f(v \pm c_2))}{\log(f(v))} \to 1$ as $|v| \to \infty$ for any $c_2 \in \Re$. Thus for any $\sigma \in (0, 1)$, there exists a constant c_3 such that, for $|v| > c_3$,

$$1 - \sigma \leq \frac{\log(f(v \pm c_2))}{\log(f(v))} \leq \frac{1}{1-\sigma}.$$

In addition, without loss of generality, we can assume $f(v) < 1$ for $|v| > c_3$ because $f(v)$ will vanish as $|v| \to \infty$. Therefore, we have

$$\log(f(v \pm c_2)) \leq (1-\sigma)\log(f(v)), \quad \text{or equivalently,} \quad f(v \pm c_2) \leq f(v)^{1-\sigma}.$$

One way to test Assumption 2(3) is to test its sufficient condition that the extreme value index (EV index) of V is positive. There is vast literature on estimating and testing EV index. We refer readers to Resnick (2007) for more detail.

Next, we state the requirement for the kernel function used to estimate $f(\cdot)$, the density of V.

Assumption 3. *Let $K(\cdot)$ denote a univariate and differentiable kernel density. $K(\cdot)$ is supported on $[-1, 1]$, is symmetric, and has order higher than ν, with bounded derivatives up to degree ν. $K(1) = K(-1) = 0$.*

Higher order kernels are commonly used in density estimation. Here in addition, we assume the kernel has a compact support, just to simplify the proof. We expect the theoretical results in the paper are still valid when using kernels that decay sufficiently fast in tails.

The key restrictions for $\sqrt{n}$-consistency of $\hat{\alpha}_n$ are on the tails of V and ε. Next, we introduce some definitions from the extreme value theory that help us characterize the tail behaviors of probability distributions.

[1] $f(\cdot)$ is regularly varying at ∞.

The cumulative distribution function (CDF) F belongs to the domain of attraction of type 1 or 2 generalized extreme value (GEV) distributions if

$$\begin{aligned}
&\text{type 1 tails } (\xi = 0): \quad \text{as } z \to +\infty \quad (1-F)(z + va(z)) \sim (1-F)(z)e^v, \quad \forall v \in \Re, \\
&\text{type 2 tails } (\xi > 0): \quad \text{as } z \to +\infty \quad (1-F)(vz) \sim v^{-1/\xi}(1-F)(z), \quad \forall v > 0,
\end{aligned}$$

in which $a(z) = \int_z^\infty (1-F)(v)dv/(1-F)(z)$ and ξ is the EV index.

In addition, we write $G \in RV_a(s)$ for some constant a, and $s = 0$ or ∞, if $\frac{G(xt)}{G(t)} \to x^a$ as $t \to s$ for any $x > 0$. The inverse of a CDF G is written as $G^{\leftarrow}(\tau) = \inf\{t : G(t) > \tau\}$. The inverse of a survival function $1 - G$ is $(1-G)^{\leftarrow}(\tau) = \inf\{t : (1-G)(t) \leq \tau\}$.[2]

Now we are ready to state the regularity conditions for the tails of V and ε.

Assumption 4. *Let F and F_ε be the CDF of V and ε, respectively.*
(1) $F(v)$ and $1 - F(-v)$ are in the attraction domain of type 1 or 2 GEV distributions with EV indices ξ_r and ξ_l, respectively.
(2) $F_\varepsilon(e)$ and $1 - F_\varepsilon(-e)$ are in the attraction domain of type 1 or 2 GEV distributions with EV indices λ_r and λ_l, respectively.

Assumption 4(1) and 4(2) are satisfied by almost all well-known continuous distributions with an unbounded support. We refer readers to Resnick (1987) for further discussion of these conditions.

Next, we turn to the relative thickness of the tails of V and ε, which has been identified by Khan and Tamer (2010) as the key condition for determining the convergence rate of semiparametric estimators of α.

Assumption 5. *For the right tail of the distribution of V and ε, One of the following three tail restrictions is satisfied, and the symmetric condition holds for the left tail.*
(1) $\xi_r > 0$ and $\lambda_r = 0$.
(2) $\xi_r > 0$, $\lambda_r > 0$ and $\frac{1}{1+\sigma} > \frac{(1+\xi_r)\lambda_r}{\xi_r(1-\lambda_r)}$.
(3) $\xi_r = 0$, $\lambda_r = 0$, $1 - F(t) = \exp(-T_r(t))$ with $T_r(t) \in RV_{d_{1,r}}(\infty)$, $1 - F_\varepsilon(t) = \exp(-D_r(t))$ with $D_r(t) \in RV_{d_{2,r}}(\infty)$, and $\infty \geq d_{2,r} > d_{1,r} \geq 0$.

Assumption 5 is the sufficient tail restriction for our estimator to be $\sqrt{n}$-consistent. Assumption 5(1) implies ε and V have rapidly varying and regularly varying tails, respectively. Assumption 5(2) considers the situation in which both ε and V have regularly varying tails. Then the tail restriction is on the relative magnitude of the two EV indices. Assumption 5(3) considers the case in which both tails are rapidly varying, in which case the varying speeds can no longer be compared using the EV indices. The condition presented here can be viewed as a restriction on the varying index of

[2]Note here that $(1-G)^{\leftarrow}(\tau) = G^{\leftarrow}(1-\tau)$ if G is continuous at $(1-G)^{\leftarrow}(\tau)$. Otherwise, $(1-G)^{\leftarrow}(\tau)$ and $G^{\leftarrow}(1-\tau)$ are not necessarily the same. Throughout the paper, we consider the case in which the special regressor V and the error term ε are both continuous random variables.

the logarithm of the tail. Overall, all three conditions imply the tails of V are thicker than the tails of ε.

In Section 1 of our supplement, we show the information of α is zero, if for any $\delta > 0$, there exists a function $C_\delta(\cdot)$ such that

$$\mathbb{E}(1 - C_\delta(\alpha + V))^2 \leq \delta \tag{3.2}$$

and

$$\mathbb{E}C_\delta(\varepsilon) = 0. \tag{3.3}$$

Let $C_\delta(t) = 1$ when $|t| < M_\delta$ for some $M_\delta \to \infty$ as $\delta \to 0$. If $C_\delta(t)$ was bounded from below when $|t| \geq M_\delta$, then (3.3) would not hold because $P(|\varepsilon| > M_\delta) \to 0$ as $M_\delta \to \infty$. In other words, we need $C_\delta(t) \to -\infty$ as $|t| \to \infty$ so that the negative part of $C_\delta(t)$ in the tails ($|t| \geq M_\delta$) can cumulate and cancel the positive part of $C_\delta(t)$ in the middle ($|t| < M_\delta$). On the other hand, (3.2) implies that $C_\delta^2(t)$ diverges to $-\infty$ slower than the decaying rate of $f_V(t)$ as $t \to \infty$. Therefore, the existence of such C_δ implies that, heuristically,

$$\begin{aligned}\text{decaying rate of } f_V(t) &> \text{diverging rate of } C_\delta^2(t) \\ &> \text{diverging rate of } C_\delta(t) = \text{ decaying rate of } f_\varepsilon(t),\end{aligned}$$

i.e., V has thinner tails than ε does. This case is ruled out by our Assumption 5. Thus, the information of α becomes positive.

Theoretically speaking, it is possible to construct a test for Assumption 5 because the CDFs of both V and ε are identified. To illustrate this, note that V_i is observable, and the identification of the CDF of V is obvious. In addition, because V is supported on the whole real line, we can identify α and the CDF of $\varepsilon - \alpha$ by

$$F_\varepsilon(\alpha + v) = \mathbb{E}(Y|V = v), \quad \forall v \in \mathbb{R}.$$

This implies that we can identify the CDF of ε.

Although the CDFs can be identified, there is no formal test for Assumption 5 in the literature. The same situation occurs in Khan and Tamer (2010), in which the support of V can be identified as $\Re$ but a formal test is lacking.

Next, we demonstrate a scenario in which we can test if Assumption 5(1) holds true. Suppose that we are ready to impose that the tails of ε behave like those of normal or logistic distributions, which are two most popular choices by applied researchers. In such cases, $\lambda_r = \lambda_l = 0$. In addition, V is continuous and observable. So we can consistently estimate ξ_r and ξ_l, following Resnick (2007). Furthermore, as argued by Lewbel, McFadden, and Linton (2011), in some experiments, the special regressor V is randomly drawn from a distribution determined by researchers, so that its EV indices are known. For example, consider an experiment where an individual is asked if he would be willing to pay more than V dollars for some product. In this case, if the EV indices for V are positive and

$\lambda_r = \lambda_l = 0$, then we know that Assumption 5(1) holds.

If the distribution of ε is unknown, in order to formally test Assumption 5, researchers have to estimate λ_r and λ_l first. In our supplement, we provide such consistent estimators. However, the proposed estimators are not precise for two reasons. First, only the information in tails is useful for estimating the tail indices, which is limited. Second, there is additional information loss when estimating λ_r and λ_l because we can only observe the binary outcome Y instead of the continuous variable ε. We believe an independent treatment is required to obtain more precise estimators of the EV indices of ε, derive their distribution theories, and construct formal tests.

Magnac and Maurin (2007) proposed a different set of tail restrictions in the same binary response model as we do. Their conditions depend upon unknown parameter values and thus are not directly linked to more conventional stochastic restrictions on heteroskedastic errors. See Chen, Khan, and Tang (2016) for more details on this point. Our tail restrictions are not nested by the tail conditions in Magnac and Maurin (2007). Conceptually, the tail restrictions in Magnac and Maurin (2007) are for point identification, whereas ours are for $\sqrt{n}$-consistency. The reason they had an identification problem is because they focused on the case in which the support of V is not the real line so that the support of ε is not necessarily nested by the support of V, especially when ε has a full support $\Re$. On the other hand, we focus on the case in which V has a full support $\Re$ so that α is point identified. To derive the $\sqrt{n}$-consistency of their estimator, Magnac and Maurin (2007) in fact relied on the support of V being compact or on the high level assumptions used in Lewbel (2000). See, for example, footnotes 9 and 10 in Magnac and Maurin (2007).

Now we can define our feasible trimming function $\hat{I}_{n,i}$ and the two end points $\hat{l}_n$ and $\hat{r}_n$. Recall that F is the CDF of the special regressor V. Ideally, we want to use $I_{n,i} = \mathbb{1}\{V_i \in [l_n, r_n]\}$ where $r_n = (1-F)^{\leftarrow}(n^{-\rho_r})$ and $l_n = F^{\leftarrow}(n^{-\rho_l})$ for some tuning parameters ρ_r and ρ_l. However, it is infeasible because F is unknown. On the other hand, since r_n and l_n are the extremal quantiles of the special regressor V, they can be estimated by order statistics. We denote the estimators for quantiles r_n and l_n by $\hat{r}_n$ and $\hat{l}_n$, respectively, in which $\hat{r}_n = V^{(n)}_{(n-m_r+1)}$ and $\hat{l}_n = V^{(n)}_{(m_l)}$. Here $m_r = \lfloor n^{1-\rho_r} \rfloor$, $m_l = \lfloor n^{1-\rho_l} \rfloor$[3], and $V^{(n)}_{(1)} \leq V^{(n)}_{(2)} \leq \cdots \leq V^{(n)}_{(n)}$ is the ascending order statistics of $\{V_i\}_{i=1}^n$. For the asymptotic properties of $\hat{r}_n$ and $\hat{l}_n$, please see Dekkers and De Haan (1989), Falk (1991), and De Haan and Ferreira (2007). Our feasible trimming function is then constructed based on $\hat{l}_n$ and $\hat{r}_n$, instead of the infeasible ones l_n and r_n.

In simulations, we show that our trimming schedule is close to trimming the density estimator $\hat{f}(v)$ by a vanishing sequence b_n, i.e., $\mathbb{1}\{\hat{f}(v) \geq b_n\}$ where b_n depends on extremal quantiles of V. This is in contrast to using a fixed b to trim the density, which is the standard approach in semiparametric estimations (e.g., Robinson, 1988). Klein and Spady (1993) did use a trimming scheme that depends on the sample size. However, their estimator is still $\sqrt{n}$-consistent even if the density is just trimmed by a fixed b. This is because their parameters of interest are regularly identified. The problem we

[3] $\lfloor a \rfloor$ is the largest integer that is smaller than a.

face is more challenging because α is identified at infinity. Using fixed b can lead to inconsistency. Even if we let b_n decay to zero, the $\sqrt{n}$-consistency is not always obtainable without tail restrictions, as has been shown in Khan and Tamer (2010).

Assumption 6. *Let $h = c_h n^{-H}$ be the tuning parameter in the kernel density estimation, in which c_h is a fixed positive constant.*
(1) $\frac{1}{2\nu} < \rho_r$ and $\frac{1+\rho_r(\xi_r+1)}{1+2\nu} < H < \min(1-(1+\sigma)\rho_r(\xi_r+1), \frac{1}{2})$. In addition, if $\xi_r > 0$, then $\rho_r > \frac{\lambda_r}{2\xi_r(1-\lambda_r)}$.
(2) $\frac{1}{2\nu} < \rho_l$ and $\frac{1+\rho_l(\xi_l+1)}{1+2\nu} < H < \min(1-(1+\sigma)\rho_l(\xi_l+1), \frac{1}{2})$. In addition, if $\xi_l > 0$, then $\rho_l > \frac{\lambda_l}{2\xi_l(1-\lambda_l)}$.

In the scenario described after Assumption 5, we can test if Assumption 5(1) holds. If it holds, then it is practically feasible to choose $\sigma = 0$ and ρ_r, ρ_l, and H that satisfy Assumption 6. If researchers are not ready to impose any knowledge on the tails of ε, then in order to choose tuning parameters, one has to estimate λ_r and λ_l. The consistent estimators of λ_r and λ_l proposed in our supplement serve this purpose. On the other hand, as mentioned earlier, it is hard to obtain a precise estimator of λ_r and λ_l, which makes the task of practically choosing our tuning parameters very difficult. Overall, how to choose the tuning parameters when estimating irregularly identified parameters is not an easy endeavor. See for example, Andrews and Schafgans (1998) and D'Haultfoeuille, Maurel, and Zhang (2016) for sample selection models; Lewbel (1997), Lewbel (2000), and Khan and Tamer (2010) for binary response models; and Khan and Tamer (2010) for treatment effect models.

Theorem 3.1. *Recall (3.1). Under Assumptions 1–6,*
(1)

$$\sqrt{n}(\hat{\alpha}_n - \alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - P_i}{f(V_i)} + o_p(1),$$

where $P_i = E(Y_i|V_i)$.
(2) Let $\Sigma = \mathbb{E}(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)})^2 - \alpha^2$, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \Gamma_{n,i}^2 - (\frac{1}{n}\sum_{i=1}^{n} \Gamma_{n,i})^2$, then $\Sigma < \infty$ and $\hat{\Sigma} \xrightarrow{p} \Sigma$.
(3) $\hat{\Sigma}^{-1/2}\sqrt{n}(\hat{\alpha}_n - \alpha) = \hat{\Sigma}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Gamma_{n,i} - \alpha) \rightsquigarrow N(0,1)$.

Several comments on the above theorem are in order. First, as discussed after Assumption 6, if ε is normal or logit and both ξ_r and ξ_l are positive, then our tail restrictions hold. In this case, our estimator is still $\sqrt{n}$-consistent and more robust than the maximal likelihood estimator (MLE) of probit or logit model, in the sense that we only exploit that the tail of the distribution of ε is normal or logit but put no restriction on the middle of the distribution. It is possible to use the difference between our estimator and the MLE to construct a specification test which is expected to have power against some $n^{-1/2}$-local alternatives.

Second, one necessary condition for the existence of $\sqrt{n}$-consistent estimator of α is that

$$\mathbb{E}\left[\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}\right]^2 = \int_{-\infty}^{+\infty} \frac{\mathbb{E}(|Y - \mathbb{1}\{V > 0\}||V = v)}{f(v)} dv < \infty.$$

Although the density $f(v)$ will vanish at $\pm\infty$, the numerator $\mathbb{E}(|Y - \mathbb{1}\{V > 0\}||V = v)$ vanishes too. If the numerator vanishes at a much faster rate than the denominator does, their ratio is still integrable at $\pm\infty$ w.r.t. v. Our tail restrictions exploit this intuition.

Third, notice as sample size increases, the order statistic of V will diverge. This implies the trimming interval will eventually become the real line. Thus, even for a case in which the tail restrictions do not hold, the proposed estimator can still be consistent.

Last, Theorem 3.1 extends the result in Khan and Tamer (2010) that $\sqrt{n}$-consistency can be obtained in the binary response model if the special regressor has an infinite variance. In fact, Assumption 5 allows for any moments of V to exist. This extension is based on (1) additional knowledge of the tail behaviors of V and ε, and (2) a careful calculation made possible by trimming based on extremal quantiles.

Next, we show that the estimator $\hat{\alpha}_n$ proposed in Theorem 3.1 is in fact asymptotically efficient and the efficient function is $\tilde{\psi} = \frac{Y - \mathbb{E}(Y|V)}{f(V)}$.

Corollary 3.1. *Under the conditions in Theorem 3.1, $\tilde{\psi}$ is the efficient function for α, and $\hat{\alpha}_n$ is asymptotically efficient with asymptotic distribution $\mathcal{N}(0, \mathbb{E}\tilde{\psi}^2)$.*

Theorem 3.1 and Corollary 3.1 show that the tail restrictions in Assumption 5 are sufficient for the regular identification of α when the support of ε is the whole real line and the proposed estimator is asymptotically efficient. The estimator is asymptotically equivalent to the estimator proposed in Lewbel (1997) when his high level assumptions for $\sqrt{n}$-consistency hold. The efficiency bound is also the same as the one derived in Magnac and Maurin (2007) and Jacho-Chávez (2009), though the underlying assumptions for $\sqrt{n}$-consistency are different. In fact, the tail restrictions in Assumption 5 do not affect the efficiency score. The same situation occurs in Magnac and Maurin (2007), as their tail symmetry condition does not affect the efficiency bound.

3.3 An Impossibility Result

Researchers may be concerned about the necessity of this type of tail restriction. Assumption 7 characterizes situations of tails in which there does not exist any regular semiparametric estimator for α. It roughly means α is not $\sqrt{n}$-estimable. Assumption 7 can be viewed as the reverse of Assumption 5 in the sense that the roles of V and ε in Assumption 5 are reversed.

Assumption 7. *ξ_r, ξ_l, λ_r, and λ_l are defined in Assumption 4. For the right tail, one of the following three tail restrictions is satisfied or symmetric conditions for the left tail hold.*
(1) $\xi_r = 0$ and $\lambda_r > 0$,

(2) $\xi_r > 0$, $\lambda_r > 0$ *and* $\lambda_r > \frac{\xi_r}{2\xi_r+1}$.
(3) $\xi_r = 0$, $\lambda_r = 0$, $1 - F(t) = \exp(-T_r(t))$ *with* $T_r(t) \in RV_{d_{1,r}}(\infty)$, $1 - F_\varepsilon(t) = \exp(-D_r(t))$ *with* $D_r(t) \in RV_{d_{2,r}}(\infty)$, *and* $\infty \geq d_{1,r} > d_{2,r} \geq 0$.

In short, Assumption 7 requires that the tails of V are thinner than the tails of ε. Section 3.1 of Khan and Tamer (2010) considered several examples of distributions of special regressor V and error term ε and found when the tail of the special regressor V is as thin or thinner than the tail of the error term, the convergence rate for the estimator of the intercept term in (2.1) is slower than the parametric rate. Theorem 3.2 extends Khan and Tamer's (2010) observations to general situations as described in Assumption 7.

Theorem 3.2. *Under Assumptions 1, 4, and 7, the asymptotic variance* $\mathbb{E}|\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}|^2$ *is infinite, and there does not exist a regular estimator of* α.

Theorem 3.2 shows for some DGPs, the α is irregularly identified. This confirms the result in Khan and Tamer (2010) that, without tail restrictions, the semiparametric efficiency bound for α as the worst-case bound is zero. It also shows that the high-level assumptions for $\sqrt{n}$-consistency in Stoker (1991) do not hold in general.

The intuition for this theorem is the same as that mentioned in Khan and Tamer (2010). By Corollary 3.1, $\frac{Y_i - P_i}{f(V_i)}$ is the efficient score when our tail restrictions hold. Suppose there exists an efficient estimator $\tilde{\alpha}_n$, then it must have the following linear expansion:

$$\sqrt{n}(\tilde{\alpha}_n - \alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - P_i}{f(V_i)} + o_p(1).$$

However, $\mathbb{E}|\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}|^2 = \mathbb{E}|\frac{Y_i - P_i}{f(V_i)}|^2 + \alpha^2 = \infty$ under Assumption 7. This implies the second moment of $\frac{Y_i - P_i}{f(V_i)}$ is infinite, and thus, the efficient estimator of α has an infinite asymptotic variance. This is a contradiction.

4 Extensions

This section extends the method of trimming by extremal quantiles to the estimator of the binary choice model proposed by Dong and Lewbel (2015) as follows:

$$Y_i = \mathbb{1}\{X_i'\beta + V_i - \varepsilon_i \geq 0\}.$$

V is our special regressor and its coefficient is normalized to one. Let $X = [X_1, Z_1]$ where X_1 and Z_1 are the endogenous and exogenous elements of X, respectively. Both X_1 and Z_1 can contain discrete elements. The error term ε can be heteroskedastic with respect to X. $Z = [Z_1, Z_2]$ is a set of instrumental variables (IVs).

Dong and Lewbel's (2015) estimator, which we will define later, is more relevant for empirical applications than the toy estimator we considered in the previous section. It allows for endogenous or mismeasured regressors and heteroskedastic errors; is numerically trivial to implement; and, unlike the control function approach, can be used with limited, censored, or discrete endogenous regressions.

However, as discussed in Dong and Lewbel (2015), the formal limiting distribution theory for their estimator is still lacking, mainly because the β's are irrgularly identified. We fill this gap by deriving a $\sqrt{n}$-consistent estimator of β under additional tail restrictions similar to those in Section 3.2. In particular, we consider the following setup adapted from Corollary 1 of Dong and Lewbel (2015).

Assumption 8. $\mathbb{E}(Z\varepsilon) = 0$, $\Sigma_{xz} = \mathbb{E}XZ'$ *has full column rank,* $V = S'\gamma + U$, $\mathbb{E}U = 0$, $U \perp (S, \varepsilon)$ *where* $S = (X, Z)$. U *has density* $f(U)$, *and its support is the real line* $\Re$.

Several remarks are in order. First, the full column rank of Σ_{xz} implies the number of IVs is greater than or equal to the number of endogenous variables, which is necessary for point identification. Second, we will assume that U has a full support, which directly implies the support condition in Corollary 1 of Dong and Lewbel (2015). Last, since $U \perp (\varepsilon, S)$, we have $V \perp \varepsilon | S$, which is Assumption A.2 in Lewbel (2000). Then, based on Theorem 1 of Dong and Lewbel (2015), β is identified as

$$\beta = \Delta \mathbb{E}\left(Z \frac{Y - \mathbb{1}\{V > 0\}}{f(V|S)} \right), \tag{4.1}$$

in which $\Delta = (\Sigma_{xz} W \Sigma_{xz}')^{-1} \Sigma_{xz} W$ and W is the usual weighting matrix. A popular choice for W is Σ_{zz}^{-1} where $\Sigma_{zz} = \mathbb{E}ZZ'$.

The identification of β does not rely on the condition that $V = S'\gamma + U$. The purpose of the latter condition is to reduce the dimensionality. Based on (4.1), when estimating $f(V|S)$ nonparametrically, we will suffer from the curse of dimensionality. Dong and Lewbel (2015) imposed a parsimonious parametric model such that $V = S'\gamma + U$. Under this parametric assumption, we need to estimate only $f(u)$, the density of U, which is univariate. In fact, β can be identified as

$$\beta = \Delta \mathbb{E}\left(Z \frac{Y - \mathbb{1}\{V > 0\}}{f(U)} \right). \tag{4.2}$$

As for the $\sqrt{n}$-consistency for the estimator of β, the intuition from the previous section still applies: the convergence rate depends upon the relative thickness between the tails of ε and U. Next, we impose sufficient tail restrictions for β to be $\sqrt{n}$-consistently estimable. Compared to the simplified model considered in Section 3.2, the additional difficulty here is that U_i is not directly observable. We propose to replace it by the residual $\hat{U}_i$ from the regression of V on S.

In order to give a formal definition of our semiparametric estimator and the trimming scheme, we need the following assumption.

Assumption 9. *The support of S is compact.*

Assumption 9 implies the tails of V and U are the same. This ensures the tail restrictions on U are sufficient for $\sqrt{n}$-consistency. If the support of Z is in fact unbounded, we can trim Z by $\tau(Z) = \mathbb{1}\{Z \in \mathcal{Z}_0\}$, where $\mathcal{Z}_0$ is a compact subset of $\Re^{d_z}$ and d_z is the dimension of Z. If $\mathbb{E}(\varepsilon Z) = 0$ is replaced by $\mathbb{E}(\varepsilon|Z) = 0$, then β can still be identified as

$$\beta = \tilde{\Delta}\mathbb{E}\left(Z\tau(Z)\frac{Y - \mathbb{1}\{V > 0\}}{f(U)}\right),$$

in which $\tilde{\Delta} = (\tilde{\Sigma}_{xz}W\tilde{\Sigma}'_{xz})^{-1}\tilde{\Sigma}_{xz}W$, $\tilde{\Sigma}_{xz} = \mathbb{E}XZ'\tau(Z)$, and W is the usual weighting matrix.

If some elements of the endogenous variable X_1 are unbounded, to trim X_1 as above does not maintain the moment equality. Fortunately, in many empirical applications, the endogenous variables are bounded, and sometimes even discrete.

Given the identification of β, we propose to estimate β in two steps. In the first step, we regress V on S and obtain the OLS estimator $\hat{\gamma}$ of γ. Then we compute the residual $\hat{U}_i$ as

$$\hat{U}_i = V_i - S'\hat{\gamma}.$$

In the second step, β is estimated as

$$\hat{\beta} = (\hat{\Sigma}_{xz}W_n\hat{\Sigma}'_{xz})^{-1}\hat{\Sigma}_{xz}W_n\hat{\Phi},$$

in which W_n is a (random) weighting matrix such that $W_n \xrightarrow{p} W$ for some positive definite (non-random) matrix W,

$$\hat{\Sigma}_{zx} = \frac{1}{n}\sum_{i=1}^{n} Z_i X_i',$$

$$\hat{\Phi} = \frac{1}{n}\sum_{i=1}^{n} Z_i\left(\frac{Y_i - \mathbb{1}\{V_i \geq 0\}}{\tilde{f}(\hat{U}_i)}\right)\tilde{I}_{n,i}, \tag{4.3}$$

$\tilde{f}(\hat{U}_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K(\frac{\hat{U}_j - \hat{U}_i}{h})$, and $\tilde{I}_{n,i} = \mathbb{1}\{\hat{U}_i \in (\tilde{l}_n, \tilde{r}_n)\}$. Here $\tilde{r}_n = \hat{U}^{(n)}_{(n-m_r+1)}$ and $\tilde{l}_n = \hat{U}^{(n)}_{(m_l)}$ in which $m_r = [n^{1-\rho_r}]$ and $m_l = [n^{1-\rho_l}]$, respectively. $\tilde{I}_{n,i}$ is the feasible estimator of $I_{n,i} = \mathbb{1}\{U_i \in S_n\}$, in which $S_n = (l_n, r_n)$, $r_n = (1-F)^{\leftarrow}(n^{-\rho_r})$, $l_n = F^{\leftarrow}(n^{-\rho_l})$, and F is the CDF of U.

Assumption 10. *Let $f(\cdot)$, $F(\cdot)$ and $F_\varepsilon(\cdot)$ denote the density of U, the CDF of U, and the CDF of ε, respectively. Then Assumptions 2 and 4 hold for $f(\cdot)$, $F(\cdot)$, and $F_\varepsilon(\cdot)$.*

Assumption 10 relies on the same tail regularity assumptions we used in the previous section. Next, we state our tail restrictions on the relative thickness of the tails between U and ε.

Assumption 11. *ξ_r, ξ_l, λ_r, and λ_l are defined in Assumption 4. For the right tail, one of the following three tail restrictions is satisfied, and the symmetric conditions hold for the left tail.*

(1) $\xi_r > 0$ *and* $\lambda_r = 0$.
(2) $\xi_r > 0$, $\lambda_r > 0$, *and* $\frac{1}{2+\sigma} > \frac{(1+\xi_r)\lambda_r}{\xi_r(1-\lambda_r)}$.
(3) $\xi_r = 0$, $\lambda_r = 0$, $1 - F(t) = \exp(-T_r(t))$ *with* $T_r(t) \in RV_{d_{1,r}}(\infty)$, $1 - F_\varepsilon(t) = \exp(-D_r(t))$ *with* $D_r(t) \in RV_{d_{2,r}}(\infty)$, *and* $\infty \geq d_{2,r} > d_{1,r} \geq 0$.

Since γ is identified, so is the CDF of $U_i = V_i - S_i'\gamma$. In addition, β is identified. Thus, as discussed after Assumption 5, the CDF of ε is also identified. However, as discussed previously, a feasible test for Assumption 11 is lacking. It would be useful for future research to construct a feasible statistical test for the tail restrictions.

On the other hand, in experiment, V is randomly generated from a distribution which is known to researchers. If the distribution has positive EV indices, then Assumption 11(1) holds, given the tails of ε behave like normal or logit. In this scenario again, the new estimator is more robust to MLE of probit or logit models since it puts no restriction on the middle of the distribution of ε.

Given Assumption 11, we choose the two key tuning parameters ρ_r and ρ_l as follows:

Assumption 12. *Let* $h = c_h n^{-H}$ *be the tuning parameter in the kernel density estimator for some positive constants* c_h *and* H.
(1) $\frac{1}{2\nu} < \rho_r$ *and* $\frac{1+\rho_r(\xi_r+1)}{1+2\nu} < H < \min(1-(1+\sigma)\rho_r(\xi_r+1), \frac{1}{4})$. *In addition, if* $\xi_r > 0$, *then* $\rho_r > \frac{\lambda_r}{2\xi_r(1-\lambda_r)}$.
(2) $\frac{1}{2\nu} < \rho_l$ *and* $\frac{1+\rho_l(\xi_l+1)}{1+2\nu} < H < \min(1-(1+\sigma)\rho_l(\xi_l+1), \frac{1}{4})$. *In addition, if* $\xi_l > 0$, *then* $\rho_l > \frac{\lambda_l}{2\xi_l(1-\lambda_l)}$.

Comparing Assumption 12 with Assumption 6, the key difference is the upper bound of H. This is due to the fact that U_i is not observed and the estimator $\hat{U}_i$ is used to compute $\tilde{f}(\cdot)$, the kernel estimator of the density of U. If U_i is observed, the accuracy of the univariate kernel density estimator is $O_p(\sqrt{\frac{L_n}{nh}})$. Now since $\hat{U}_i$ is used, the accuracy becomes $O_p(\frac{1}{\sqrt{nh^4}})$.

The next theorem establishes the $\sqrt{n}$-consistency of $\hat{\beta}$.

Theorem 4.1. *If* $W_n \xrightarrow{p} W$ *for some positive definite (nonrandom) matrix* W, *and if Assumptions 3 and 8–12 hold, then*

$$E|\Psi_i|^{2+\sigma} < \infty$$

and

$$\sqrt{n}(\hat{\beta} - \beta) = (\Sigma_{xz} W \Sigma_{xz}')^{-1} \Sigma_{xz} W \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Psi_i - (Z_i X_i' - \Sigma_{zx})\beta) + o_p(1) \rightsquigarrow \mathcal{N}(0, \Sigma_\beta),$$

in which

$$\Psi_i = \frac{Z_i(Y_i - \mathbb{1}\{V_i > 0\})}{f(U_i)} - \frac{\mathbb{E}(Z_i(Y_i - \mathbb{1}\{V_i > 0\})|U_i)}{f(U_i)} + \mathbb{E}\left(\frac{Z_i(Y_i - \mathbb{1}\{V_i > 0\})f'(U_i)(S_i - ES_i)'}{f(U_i)^2}\right)\Sigma_{ss}^{-1} S_i U_i,$$

$$\Sigma_\beta = (\Sigma_{zx}' W \Sigma_{zx})^{-1} \Sigma_{zx}' W \Sigma_0 W \Sigma_{zx} (\Sigma_{zx}' W \Sigma_{zx})^{-1},$$

$$\Sigma_0 = \mathbb{E}(\Psi_i - (Z_i X_i' - \Sigma_{zx})\beta)(\Psi_i - (Z_i X_i' - \Sigma_{zx})\beta)',$$

and $\Sigma_{ss} = \mathbb{E}SS'$.

5 Simulations

In this section, we exploit the finite sample performance of our estimator with trimming by extremal quantiles. We use the same model as in the simulation section of Lewbel (1997):

$$Y_i = \mathbb{1}\{-1 + V_i - \varepsilon_i > 0\}.$$

We consider six simulation designs. The first five designs satisfy Assumption 5 so that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal. The last design satisfies Assumption 7 so that no regular estimator exists. All simulations are repeated for 1,000 replications and the sample sizes considered are 200, 400, 800, 1600, 3200, and 6400.

For each design, we report the bias and root-mean-square error (rMSE) of our estimator (denoted as "Ex") and four estimators computed by trimming out the observations for which the estimated density is too small. In particular, these four estimators take the form of

$$\hat{\alpha}_L(b) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i - \mathbb{1}\{V_i > 0\}}{\hat{f}(V_i)} \mathbb{1}\{\hat{f}(V_i) > b\}$$

with $b \in (0.05, 0.01, 0.002, 0)$. The corresponding estimators are named "L1–L4." In particular, L4 is the untrimmed estimator.

Although Lewbel (1997) suggested choosing b such that it converges to zero as $n \to \infty$, the author did not investigate the finite sample performance nor provide any suggestion on such trimming constant. Here, we propose such a constant based on extremal quantiles and denote it as $\hat{b}$, where

$$\hat{b} = \min(\hat{f}(\hat{l}_n), \hat{f}(\hat{r}_n)).$$

Then, $\hat{\alpha}_L(\hat{b})$, the estimator corresponding to $\hat{b}$, is denoted as "L" and its finite sample performances are reported for each design.

To estimate the density, we use the fourth-order Epanechnikov kernel;[4] that is,

$$k(u) = \frac{45}{32}(1 - \frac{7}{3}u^2)(1 - u^2)\mathbb{1}\{|u| < 1\}.$$

The tuning parameter for the kernel density estimation takes the form of $h = c_h n^{-H}$ where c_h is a

[4] This implies $\nu = 4$.

positive constant and H is defined in Assumption 6. Since the bias is of order of h^4, based on Powell and Stoker (1996) and the results of numerical integration in Lewbel (1997), we set the constant as

$$c_h = \left(\frac{2.532^s \times 8}{0.0204 \times 2 \times s} \right)^{\frac{1}{8+s}},$$

in which s is the dimension of V and $s = 1$ in our simulation.

For H, when computing $\hat{\alpha}_L(b)$ with $b \in (0.05, 0.01, 0.002, 0)$, i.e., "L1–L4", we use the optimal rate $H^* = \frac{2}{s+8}$. Our H^* is different from that proposed in Lewbel (1997) because we use a fourth-order kernel while Lewbel (1997) used a second-order one.

When computing $\hat{\alpha}_n$ and $\hat{\alpha}_L(\hat{b})$, i.e., "Ex" and "L", we will specify the tuning parameters and verify Assumption 6 case by case.

Design 1

$V \sim T(6)$ and $\varepsilon \sim \mathcal{N}(0,1)$. The EV indices for V are $\xi_l = \xi_r = \frac{1}{6}$, $\sigma = 0$, and $\lambda_l = \lambda_r = 0$ for ε. This implies that Assumption 4 holds. By choosing $\rho_r = \rho_l = \frac{1.9}{3}$, Assumption 6(1) implies

$$\frac{31.3}{162} < H < \frac{4.7}{18}.$$

Since the optimal rate $H^* = \frac{2}{9}$ satisfies the above inequality, we choose $H = H^*$ when computing $\hat{\alpha}_n$ and $\hat{\alpha}_L(\hat{b})$. Table 1 shows the biases and rMSEs of the estimators.

	Bias						Root-MSE					
N	Ex	L	L1	L2	L3	L4	Ex	L	L1	L2	L3	L4
200	0.084	0.062	0.071	0.017	0.016	0.016	0.150	0.145	0.145	0.163	0.165	0.165
400	0.045	0.034	0.067	0.016	0.013	0.013	0.107	0.108	0.113	0.111	0.115	0.115
800	0.019	0.013	0.060	0.006	0.002	0.002	0.073	0.075	0.087	0.079	0.084	0.084
1,600	0.011	0.010	0.062	0.010	0.006	0.006	0.055	0.055	0.077	0.055	0.057	0.057
3,200	0.004	0.004	0.060	0.006	0.002	0.002	0.041	0.041	0.068	0.040	0.042	0.043
6,400	0.001	0.001	0.059	0.004	0.000	0.000	0.029	0.029	0.064	0.028	0.030	0.030

Table 1: Biases and rMSEs

Since the rMSEs for Ex decrease at rate $\sqrt{2}$ as sample size doubles, Ex is $\sqrt{n}$-consistent. This provides evidence that even when V has a finite second moment, α is still $\sqrt{n}$-estimable. L has similar performance as Ex. In addition, when $b = 0.05$, the estimator L1 has non-vanishing biases. This is not surprising because the threshold b does not vanish as sample size increases. When we choose a smaller b, the biases for estimators L2 and L3 are smaller. However, they are still fixed. So asymptotically, the biases cannot vanish, although for the current sample sizes considered, the biases are relatively small compared to the estimation errors. Last, our estimators Ex and L perform as well as L2–L4 in terms of rMSEs.

Design 2

Next, we consider the case in which ε is not symmetrically distributed. In particular, we set $V \sim T(6)$ and $\varepsilon = \frac{e_1+e_2^2+e_3^2-2}{\sqrt{5}}$ where (e_1, e_2, e_3) are independent standard normals. As in the first design, $\xi_l = \xi_r = \frac{1}{6}$, $\sigma = 0$, and $\lambda_l = \lambda_r = 0$. This implies that Assumption 5(1) holds and $\hat{\alpha}_n$ is $\sqrt{n}$-consistent. We choose the same set of tuning parameters as in Design 1; that is,

$$\rho_r = \rho_l = \frac{1.9}{3}, \quad H = H^*.$$

	Bias						Root-MSE					
N	Ex	L	L1	L2	L3	L4	Ex	L	L1	L2	L3	L4
200	0.097	0.078	0.103	0.037	0.035	0.035	0.160	0.157	0.159	0.175	0.175	0.175
400	0.057	0.045	0.099	0.030	0.022	0.022	0.113	0.112	0.129	0.115	0.126	0.126
800	0.041	0.032	0.102	0.031	0.019	0.019	0.087	0.087	0.118	0.085	0.097	0.097
1,600	0.021	0.016	0.094	0.024	0.008	0.008	0.066	0.068	0.103	0.064	0.076	0.077
3,200	0.011	0.008	0.095	0.021	0.004	0.003	0.050	0.051	0.100	0.048	0.053	0.055
6,400	0.008	0.006	0.095	0.022	0.006	0.005	0.038	0.038	0.098	0.037	0.038	0.040

Table 2: Biases and rMSEs

We see from Table 2 that estimator Ex is indeed $\sqrt{n}$-consistent and both estimators Ex and L perform better than the un-truncated estimator L4 in terms of rMSEs.

Design 3

Here we consider the case in which V is not symmetrically distributed and the tails of ε decay polynomially. In particular,

$$V = T_1 + T_2^2 - 2, \quad V \sim T(11)$$

where $T_1 \sim T(6)$, $T_2 \sim T(4)$, and $T_1 \perp\!\!\!\perp T_2$. We have $\xi_r = \frac{1}{2}$, $\xi_l = \frac{1}{6}$, $\sigma = 0$, and $\lambda_r = \lambda_l = \frac{1}{11}$.

It is easy to check that Assumption 5(2) holds. Therefore, $\hat{\alpha}_n$ is $\sqrt{n}$-consistent. Choose $\rho_r = \frac{1}{2}$ and $\rho_l = \frac{1.9}{3}$. This implies that

$$\frac{7}{36} < H < \frac{1}{4} \quad \text{and} \quad \frac{31.3}{162} < H < \frac{4.7}{18}.$$

$H = H^*$ suffices.

	Bias						Root-MSE					
N	Ex	L	L1	L2	L3	L4	Ex	L	L1	L2	L3	L4
200	0.062	0.072	0.240	0.056	0.055	0.055	0.228	0.229	0.310	0.228	0.230	0.230
400	0.032	0.035	0.214	0.034	0.033	0.033	0.155	0.155	0.257	0.156	0.156	0.156
800	0.017	0.018	0.197	0.019	0.018	0.018	0.113	0.113	0.224	0.113	0.113	0.113
1,600	0.011	0.011	0.187	0.012	0.011	0.011	0.077	0.077	0.203	0.077	0.077	0.077
3,200	0.003	0.003	0.177	0.004	0.003	0.003	0.054	0.054	0.187	0.054	0.054	0.054
6,400	0.005	0.005	0.178	0.005	0.005	0.005	0.039	0.040	0.184	0.039	0.039	0.040

Table 3: Biases and rMSEs

From Table 3, we note that Ex and L are $\sqrt{n}$-consistent and our estimation method performs well under asymmetry tail behaviors of V.

Design 4

For the fourth design, we consider the case in which V has exponentially decaying tails. In particular, $V \sim \mathcal{N}(0,1)$, $\varepsilon = \text{sign}(e_1)|e_1|^{\frac{1}{3}}$, and $e_1 \sim \mathcal{N}(0,1)$. In this case, $\xi_r = \xi_l = \lambda_r = \lambda_l = 0$, but Assumption 5(3) holds because $d_{1,r} = d_{1,l} = 2$ and $d_{2,r} = d_{2,l} = 6$. By Assumption 6(1), we can set $\rho_r = \rho_l = \frac{1}{4}$ and $H = H^*$.

	Bias						Root-MSE					
N	Ex	L	L1	L2	L3	L4	Ex	L	L1	L2	L3	L4
200	0.105	0.077	0.041	0.018	0.018	0.018	0.152	0.139	0.127	0.131	0.131	0.131
400	0.040	0.029	0.027	0.006	0.006	0.006	0.095	0.092	0.091	0.093	0.093	0.093
800	0.017	0.014	0.028	0.007	0.007	0.007	0.064	0.063	0.068	0.064	0.064	0.064
1,600	0.005	0.005	0.022	0.004	0.004	0.004	0.045	0.046	0.048	0.046	0.046	0.046
3,200	0.001	0.001	0.018	0.001	0.001	0.001	0.032	0.032	0.036	0.032	0.032	0.032
6,400	0.001	0.001	0.019	0.001	0.001	0.001	0.022	0.022	0.028	0.022	0.022	0.022

Table 4: Biases and rMSEs

From Table 4, we see that even though every moment of V exists, Ex and L can still be $\sqrt{n}$-consistent.

Design 5

In this design, we consider the case in which both V and ε have exponentially decaying tails and the CDF of ε is not symmetric. In particular, $V = e_1^3$, $\varepsilon = \frac{e_2+e_3^2+e_4^2-2}{\sqrt{5}}$ where e_1, e_2 ,e_3, and e_4 are standard normally distributed and mutually independent. In this case, $\xi_r = \xi_l = \lambda_r = \lambda_l = 0$, but Assumption 5(3) holds. We can choose ρ_r, ρ_l, and H as we did in Design 4; that is, $\rho_r = \rho_l = \frac{1}{4}$ and $H = H^*$.

	Bias						Root-MSE					
N	Ex	L	L1	L2	L3	L4	Ex	L	L1	L2	L3	L4
200	0.083	0.082	0.257	0.079	0.077	0.077	0.230	0.229	0.297	0.231	0.232	0.232
400	0.053	0.054	0.240	0.060	0.052	0.052	0.172	0.170	0.264	0.165	0.172	0.172
800	0.036	0.036	0.228	0.044	0.036	0.036	0.120	0.119	0.241	0.116	0.120	0.120
1,600	0.036	0.036	0.227	0.042	0.036	0.036	0.082	0.082	0.234	0.083	0.082	0.082
3,200	0.023	0.023	0.219	0.028	0.023	0.023	0.061	0.061	0.224	0.061	0.061	0.061
6,400	0.019	0.019	0.215	0.024	0.019	0.019	0.045	0.045	0.218	0.046	0.045	0.045

Table 5: Biases and rMSEs

From Table 5, we see that when b is relatively large, as in L1, the biases do not vanish. When b is small, the estimators L2-L4 have similar performances as our estimators Ex and L.

Design 6

Last, we consider the case in which Assumption 5 does not hold. In particular, $V \sim T(6)$ and $\varepsilon \sim T(2)$. This implies that $\xi_r = \xi_l = \frac{1}{6}$, $\lambda_r = \lambda_l = \frac{1}{2}$, and Assumption 7(2) holds. Then, based on Theorem 3.2, there does not exist any regular estimator for α. When computing Ex and L, we choose $\rho_r = \rho_l = \frac{1.9}{3}$ and $H = H^*$, as before.

	Bias						Root-MSE					
N	Ex	L	L1	L2	L3	L4	Ex	L	L1	L2	L3	L4
200	0.216	0.180	0.194	0.102	0.102	0.102	0.263	0.244	0.246	0.269	0.269	0.269
400	0.156	0.129	0.181	0.091	0.068	0.068	0.194	0.182	0.210	0.187	0.214	0.214
800	0.135	0.117	0.187	0.098	0.069	0.069	0.166	0.156	0.203	0.149	0.188	0.188
1,600	0.108	0.094	0.183	0.095	0.053	0.052	0.134	0.128	0.190	0.127	0.155	0.157
3,200	0.079	0.067	0.181	0.084	0.049	0.040	0.102	0.097	0.186	0.103	0.108	0.137
6,400	0.069	0.060	0.182	0.087	0.050	0.037	0.089	0.085	0.184	0.097	0.087	0.116

Table 6: Biases and rMSEs

From Table 6, we first see that no estimator is $\sqrt{n}$-consistent. Second, the biases for our estimators Ex and L are still decreasing, while the biases for L1 and L2 do not vanish. The un-truncated estimator L4 achieves the smallest bias in a cost of large variance. That is why its rMSEs are much larger than those of our estimators Ex and L.

Summary

In general, we obtained three notable findings from this simulation study. First, when Assumption 5 holds, both Ex and L are $\sqrt{n}$-consistent and they have similar finite sample performances. In this case, Lewbel's (1997) estimators based on a fixed trimming constant has non-vanishing and dominant biases when the constant is not sufficiently small. When the constant is sufficiently small, Lewbel's (1997) estimators has small but fixed bias. Given the sample size we considered, Lewbel's (1997) estimators' finite sample performances are similar to those of ours. Neither our estimators nor Lewbel's (1997) dominate each other because, ignoring bias, both estimators are efficient. Second,

we verify that even if V has all its moments exist, the existence of a $\sqrt{n}$-consistency estimator of α is still possible. Last, when Assumption 7 holds, neither our estimators nor Lewbel's (1997) is $\sqrt{n}$-consistent. However, our estimators are still consistent and have smaller rMSEs than those of the un-truncated estimator. More simulation results about the inference results using t-statistics for the six designs can be found in the supplement.

6 Conclusion

Because the intercept of the binary response model is irregularly identified, the convergence rate for its semiparametric estimator depends upon the tail behaviors of the special regressor V and the unobservable ε. This paper proposes a set of primitive tail restrictions that guarantee the existence of a $\sqrt{n}$-consistent estimator of the intercept. In addition, we provided a set of opposite primitive tail restrictions under which there does not exist any regular estimator for the intercept. Given the tail restrictions for $\sqrt{n}$-consistency, we proposed a semiparametric estimator for the intercept by trimming based on extremal quantiles of the special regressor, and showed that the estimator is efficient. Last, we extended the method of trimming by extremal quantiles to allow for endogenous covariates X.

References

Andrews, D. W. and Schafgans, M. M. (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 65, 497–517.

Chaudhuri, S. and Hill, J. B. (2015), "Robust Estimation for Average Treatment Effects," *Dept. of Economics, University of North Carolina-Chapel Hill*.

Chen, S., Khan, S., and Tang, X. (2016), "Informational Content of Special Regressors in Heteroskedastic Binary Response Models," *The Journal of Econometrics*, 193, 162–182.

De Haan, L. and Ferreira, A. (2007), *Extreme Value Theory: An Introduction*, Springer Science & Business Media.

Dekkers, A. L. and De Haan, L. (1989), "On the Estimation of the Extreme-value Index and Large Quantile Estimation," *The Annals of Statistics*, 17, 1795–1832.

D'Haultfoeuille, X., Maurel, A., and Zhang, Y. (2016), "Extremal Quantile Regressions for Selection Models and the Black-White Wage Gap," Working Paper.

Dong, Y. and Lewbel, A. (2015), "A Simple Estimator for Binary Choice Models with Endogenous Regressors," *Econometric Reviews*, 34, 82–105.

Falk, M. (1991), "A Note on the Inverse Bootstrap Process for Large Quantiles," *Stochastic processes and their applications*, 38, 359–363.

Han, A. K. (1987), "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.

Hill, J. B. and Renault, E. (2010), "Generalized Method of Moments with Tail Trimming," *Dept. of Economics, University of North Carolina-Chapel Hill*.

Horowitz, J. L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-index Models," *Journal of Econometrics*, 58, 71–120.

Jacho-Chávez, D. T. (2009), "Efficiency Bounds for Semiparametric Estimation of Inverse Conditional-density-weighted Functions," *Econometric Theory*, 25, 847–855.

Khan, S. and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

Klein, R. W. and Spady, R. H. (1993), "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Lewbel, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory*, 13, 32–51.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145–177.

Lewbel, A. and Schennach, S. M. (2007), "A Simple Ordered Data Estimator for Inverse Density Weighted Expectations," *Journal of Econometrics*, 136, 189–211.

Lewbel, A., McFadden, D., and Linton, O. (2011), "Estimating features of a distribution from binomial data," *Journal of Econometrics*, 162, 170–188.

Magnac, T. and Maurin, E. (2007), "Identification and Information in Monotone Binary Models," *Journal of Econometrics*, 139, 76–104.

Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of econometrics*, 3, 205–228.

Manski, C. F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of econometrics*, 27, 313–333.

Manski, C. F. (1988), "Identification of Binary Response Models," *Journal of the American statistical Association*, 83, 729–738.

Powell, J. L. and Stoker, T. M. (1996), "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics*, 75, 291–316.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Resnick, S. I. (1987), *Extreme Values, Regular Variation, and Point Processes*, Springer.

Resnick, S. I. (2007), *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*, Springer Science & Business Media.

Robinson, P. M. (1988), "Root-N-consistent semiparametric regression," *Econometrica*, 56, 931–954.

Stoker, T. M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," *In W. Barnett, G. Tauchen and J. Powell (eds.), Nonparametric and semiparametric methods in econometrics and statistics*, pp. 99–118.

Van der Vaart, A. W. (1998), *Asymptotic Statistics*, vol. 3, Cambridge University Press.

Yang, T. (2015), "Asymptotic Trimming and Rate Adaptive Inference for Endogenous Selection Estimates," *Working Paper, The Australian National University.*

7 Appendix

This section contains the proof of Theorem 3.1, Corollary 3.1, and Theorem 3.2. The proof of Theorem 4.1 and all lemmas are collected in a supplement.

7.1 Notations

Throughout the Appendix, we denote C as a generic positive constant whose value differs in different contexts. L_n is a generic function of n, which is slowly varying as $n \to \infty$, i.e., $\frac{L_{kn}}{L_n} \to 1$ as $n \to \infty$ for any $k > 0$.

7.2 Proof of Theorem 3.1

For part (1), we first decompose $\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mathbb{1}\{V_i>0\}}{\hat{f}(V_i)}\hat{I}_{n,i}$ as follows:

$$\begin{aligned}\hat{\alpha}_n =& \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mathbb{1}\{V_i>0\}}{\hat{f}(V_i)}\hat{I}_{n,i} \\ =& \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)} I_{n,i} + \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)}\frac{f(V_i)-\hat{f}(V_i)}{f(V_i)} I_{n,i} + R_{n,1} + R_{n,2} + R_{n,3},\end{aligned} \tag{7.1}$$

in which

$$R_{n,1} = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)}\right)(\hat{I}_{n,i} - I_{n,i}),$$

$$R_{n,2} = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)}\right)\left(\frac{f(V_i)-\hat{f}(V_i)}{\hat{f}(V_i)}\right)(\hat{I}_{n,i} - I_{n,i}),$$

and

$$R_{n,3} = \frac{1}{n}\sum_{i=1}^n \left[\frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)}\right]\left[\frac{(f(V_i)-\hat{f}(V_i))^2}{f(V_i)\hat{f}(V_i)}\right] I_{n,i}.$$

By Lemma 7.1, the remainder terms are all asymptotically negligible, i.e.,

$$R_{n,1} + R_{n,2} + R_{n,3} = o_p(\frac{1}{\sqrt{n}}).$$

Hence, following (7.1), we have

$$\begin{aligned}\hat{\alpha}_n &= \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)} I_{n,i} + \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)}\frac{f(V_i)-\hat{f}(V_i)}{f(V_i)} I_{n,i} + o_p(\frac{1}{\sqrt{n}}) \\ &= \tilde{\delta}_{n,1} + \tilde{\delta}_{n,2} + o_p(\frac{1}{\sqrt{n}}).\end{aligned} \tag{7.2}$$

In (7.2), $\tilde{\delta}_{n,2}$ represents the first-order error of the first stage kernel density estimation. Next we consider the U-decomposition of $\tilde{\delta}_{n,2}$. Note

$$\tilde{\delta}_{n,2} = (C_n^2)^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P_n(W_i, W_j),$$

in which

$$\begin{aligned}P_n(W_i,W_j) = \frac{1}{2}\Bigg[&\frac{Y_i - \mathbb{1}\{V_i>0\}}{f(V_i)^2}\left(f(V_i) - \frac{1}{h}K(\frac{V_i - V_j}{h})\right) I_{n,i} \\ &+ \frac{Y_j - \mathbb{1}\{V_j>0\}}{f(V_j)^2}\left(f(V_j) - \frac{1}{h}K(\frac{V_j - V_i}{h})\right) I_{n,j}\Bigg].\end{aligned}$$

By Lemma 7.2,

$$\tilde{\delta}_{n,2} = -\frac{1}{n}\sum_{i=1}^{n} \frac{P(V_i) - \mathbb{1}\{V_i > 0\}}{f(V_i)} I_{n,i} + o_p(\frac{1}{\sqrt{n}}). \tag{7.3}$$

Combining (7.2) and (7.3), we have

$$\sqrt{n}(\hat{\alpha}_n - \alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{Y_i - P_i}{f(V_i)} I_{n,i} + o_p(1). \tag{7.4}$$

In addition, we notice $\mathbb{E}[\frac{Y_i - P_i}{f(V_i)}(1 - I_{n,i})]^2 \to 0$. This implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{Y_i - P_i}{f(V_i)} I_{n,i} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{Y_i - P_i}{f(V_i)} + o_p(1). \tag{7.5}$$

Combining (7.4) and (7.5), we obtain

$$\sqrt{n}(\hat{\alpha}_n - \alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{Y_i - P_i}{f(V_i)} + o_p(1).$$

For part (2), we have

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{\hat{f}(V_i)}\right)^2 \hat{I}_{n,i} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}\right)^2 I_{n,i} + \sum_{j=1}^{4} T_{n,j}$$

where

$$T_{n,1} = \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{\hat{f}(V_i)}\right)^2 (\hat{I}_{n,i} - I_{n,i}) I_{n,i},$$

$$T_{n,2} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{\hat{f}(V_i)}\right)^2 \left(\frac{f(V_i) - \hat{f}(V_i)}{\hat{f}(V_i)}\right) \hat{I}_{n,i},$$

$$T_{n,3} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}\right)^2 (\hat{I}_{n,i} - I_{n,i})^2,$$

and

$$T_{n,4} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}\right)^2 \left(\frac{f(V_i) - \hat{f}(V_i)}{\hat{f}(V_i)}\right)^2 \hat{I}_{n,i}.$$

Lemma 7.3 shows $T_{n,j} = o_p(1)$, for $j = 1, \cdots, 4$. This implies the desired result. Part (3) is just a combination of parts (1) and (2).

7.3 Proof of Corollary 3.1

Denote ϕ and f as the density of ε and V with a dominating measure μ, respectively. The model P_λ is indexed by parameters $\lambda = (\alpha, \phi^{\frac{1}{2}}, f^{\frac{1}{2}})$. The parameter space of λ is $\Lambda \subset \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space with inner product

$$\left\langle (\alpha, \phi^{\frac{1}{2}}, f^{\frac{1}{2}}), (\alpha', \phi'^{\frac{1}{2}}, f'^{\frac{1}{2}}) \right\rangle_{\mathcal{H}} = \alpha\alpha' + \left\langle \phi^{\frac{1}{2}}, \phi'^{\frac{1}{2}} \right\rangle_{\mathcal{L}^2(\mu)} + \left\langle f^{\frac{1}{2}}, f'^{\frac{1}{2}} \right\rangle_{\mathcal{L}^2(\mu)}$$

and some dominating measure μ.[5]

Define a functional ψ that maps the model P_λ into α, i.e., $\alpha = \psi(P_\lambda)$. Based on Lemma 25.23 in Van der Vaart (1998), in order to prove the corollary, we need to verify three conditions: (1) $\psi(P_\lambda)$ is differentiable at P_λ relative to the tangent cone $\dot{\mathcal{P}}_{P_\lambda}$ in which P_λ satisfies the tail restrictions; (2) $\tilde{\psi}$ is the efficient score; and (3)

$$\sqrt{n}(T_n - \psi(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_i + o_p(1).$$

Among them, (3) has been proved in Theorem 3.1. Next, we focus on (1) and (2).

We consider a one-parameter submodel $\lambda_t = (\alpha + th, \phi_t^{\frac{1}{2}}, f_t^{\frac{1}{2}})$ and characterize the tangent sets of α, ϕ, and f as follows. First, ${}_\alpha\dot{\mathcal{P}}_{P_{\lambda_t}}$, the tangent set of α, is $\{h \in \Re\}$. To characterize the tangent set of ϕ, we first note that η, the score of the submodel $t \to \phi_t$, is defined to satisfy the following equation:

$$\lim_{t \to 0} \int [\frac{\phi_t^{\frac{1}{2}} - \phi^{\frac{1}{2}}}{t} - \frac{1}{2}\eta\phi^{\frac{1}{2}}]^2 d\mu = 0.$$

Since $\mathbb{E}\varepsilon = 0$, the submodel should also satisfy $\int \varepsilon\phi_t(\varepsilon)d\varepsilon = 0$. Then

$$\begin{aligned}
\mathbb{E}\varepsilon\eta(\varepsilon) &= \langle \varepsilon, \eta(\varepsilon) \rangle_{\mathcal{L}^2(\phi)} \\
&= \langle 2\varepsilon\phi^{\frac{1}{2}}, \frac{1}{2}\eta\phi^{\frac{1}{2}} \rangle_{\mathcal{L}^2(d\mu)} \\
&= \lim_{t \to 0} \langle \varepsilon(\phi_t^{\frac{1}{2}} + \phi^{\frac{1}{2}}), \frac{\phi_t^{\frac{1}{2}} - \phi^{\frac{1}{2}}}{t} \rangle_{\mathcal{L}^2(d\mu)} \\
&= \lim_{t \to 0} (\int \varepsilon\phi_t(\varepsilon)d\varepsilon - \int \varepsilon\phi(\varepsilon)d\varepsilon) \\
&= 0.
\end{aligned}$$

Similarly, we can show $\mathbb{E}\eta(\varepsilon) = 0$. Thus, ${}_\phi\dot{\mathcal{P}}_{P_{\lambda_t}}$, the tangent set of ϕ, is

$$\{\eta \in \mathcal{L}^2(\phi) : \mathbb{E}\eta(\varepsilon) = 0, \mathbb{E}\varepsilon\eta(\varepsilon) = 0\}.$$

[5] Since ε and V are both assumed to be continuous random variables, μ is just the Lebesgue measure.

It is worthwhile to note the tail restrictions cannot affect the tangent set. To see this, note that any η, which is continuous on a compact support and satisfies $\mathbb{E}\eta(\varepsilon) = 0$ as well as $\mathbb{E}\varepsilon\eta(\varepsilon) = 0$, is the score function for the submodel $t \to \phi_t = (1 + t\eta(\varepsilon))\phi(\varepsilon)$. Because $\eta(\varepsilon)$ has a compact support, ϕ_t has the same tail behavior as ϕ. So the submodel satisfies the additional tail restrictions. In addition, continuous functions with compact supports are dense in $\mathcal{L}^2(\phi)$, so the (closure of) tangent set is indeed ${}_\phi\dot{\mathcal{P}}_{P_{\lambda_t}}$.

Similarly, let $g(v)$ denote the score of f_t. Then the tangent set ${}_f\dot{\mathcal{P}}_{P_{\lambda_t}}$ of f is

$$\{g \in \mathcal{L}^2(f) : \mathbb{E}g(V) = 0\}.$$

We equip ${}_\alpha\dot{\mathcal{P}}_{P_{\lambda_t}} \times_\phi \dot{\mathcal{P}}_{P_{\lambda_t}} \times_f \dot{\mathcal{P}}_{P_{\lambda_t}}$ with the inner product

$$\langle (h, \eta, g), (h', \eta', g') \rangle = hh' + \langle \eta, \eta' \rangle_{\mathcal{L}^2(P_\lambda)} + \langle g, g' \rangle_{\mathcal{L}^2(P_\lambda)}.^6$$

Let A: ${}_\alpha\dot{\mathcal{P}}_{P_{\lambda_t}} \times_\phi \dot{\mathcal{P}}_{P_{\lambda_t}} \times_f \dot{\mathcal{P}}_{P_{\lambda_t}} \to \mathcal{L}^2(P_\lambda)$ be the score operator that maps the score of λ_t (functions of (V, ε)) to the score of P_{λ_t} (the function of (Y, V)), and Φ be the CDF of ε. Model P_λ has log likelihood

$$y \log(\Phi(\alpha + v)) + (1 - y) \log(1 - \Phi(\alpha + v)) + \log(f(v)). \tag{7.6}$$

Then, by taking the ordinary derivatives of the log likelihood in (7.6) w.r.t. t, we obtain

$$A(h, \eta, g) = \dot{l}_\alpha h + \dot{l}_\phi \eta + \dot{l}_f g$$

where

$$\dot{l}_\alpha = Y \frac{\phi(V + \alpha)}{\Phi(V + \alpha)} - (1 - Y) \frac{\phi(V + \alpha)}{1 - \Phi(V + \alpha)},$$
$$\dot{l}_\phi \eta(Y, V) = \mathbb{E}(\eta(\varepsilon)|Y, V),$$

and

$$\dot{l}_f g(Y, V) = \mathbb{E}(g(V)|Y, V).$$

Let $A^* : \mathcal{L}^2(P_\lambda) \to_\alpha \dot{\mathcal{P}}_{P_{\lambda_t}} \times_\phi \dot{\mathcal{P}}_{P_{\lambda_t}} \times_f \dot{\mathcal{P}}_{P_{\lambda_t}}$ be the adjoint of A. Then, by the definition of adjoint, for any $b \in \mathcal{L}^2(P_\lambda)$,

$$\begin{aligned} \langle (h, \eta, g), A^* b \rangle &= \langle A(h, \eta, g), b \rangle = \langle \dot{l}_\alpha h + \dot{l}_\phi \eta + \dot{l}_f g, b \rangle \\ &= \langle \dot{l}_\alpha h, b \rangle + \langle \dot{l}_\phi \eta, b \rangle + \langle \dot{l}_f g, b \rangle \\ &= \langle h, \langle \dot{l}_\alpha, b \rangle_{\mathcal{L}^2(P_\lambda)} \rangle + \langle \eta, \dot{l}_\phi^* b \rangle + \langle g, \dot{l}_f^* b \rangle \\ &= \langle (h, \eta, g), (\langle \dot{l}_\alpha, b \rangle_{\mathcal{L}^2(P_\lambda)}, \dot{l}_\phi^* b, \dot{l}_f^* b) \rangle. \end{aligned} \tag{7.7}$$

[6] $\mathcal{L}^2(P_\lambda)$ means $\mathcal{L}^2$ norm w.r.t. probability P_λ.

Therefore,

$$A^*b = (\langle \dot{l}_\alpha, b\rangle_{\mathcal{L}^2(P_\lambda)}, \dot{l}_\phi^* b, \dot{l}_f^* b).$$

By Lemma 25.34 in Van der Vaart (1998) with $X = (Y, V)$ and $Y = (V, \varepsilon)$, we have

$$\dot{l}_\phi^* b(V, \varepsilon) = \mathbb{E}(b(Y, V)|V, \varepsilon)$$

and

$$\dot{l}_f^* b(V, \varepsilon) = \mathbb{E}(b(Y, V)|V, \varepsilon).$$

By Theorem 25.31 of Van der Vaart (1998), $\psi(P_\lambda)$ is differentiable at P_λ relative to $\dot{\mathcal{P}}_{P_\lambda}$ if and only if there exists $\tilde{\psi} \in \mathcal{L}^2(P_\lambda)$ such that $\langle A^*\tilde{\psi}, (h, \eta, g)\rangle = h$.[7] Such $\tilde{\psi}$ is called the efficient score.

We claim that $\tilde{\psi} = \frac{Y - \mathbb{E}(Y|V)}{f(V)}$ satisfies all the requirements above. First, Theorem 3.1 has shown, under the tail restrictions, $\tilde{\psi} \in \mathcal{L}^2(P_\lambda)$. In addition,

$$\langle \tilde{\psi}, \dot{l}_\alpha\rangle_{\mathcal{L}^2(P_\lambda)} = 1,$$

$$\langle \dot{l}_\phi^* \tilde{\psi}, \eta\rangle_{\mathcal{L}^2(P_\lambda)} = \mathbb{E}[\mathbb{E}(\tilde{\psi}|\varepsilon)\eta(\varepsilon)] = \mathbb{E}\varepsilon\eta(\varepsilon) = 0,$$

and

$$\langle \dot{l}_\phi^* \tilde{\psi}, \eta\rangle_{\mathcal{L}^2(P_\lambda)} = \mathbb{E}[\mathbb{E}(\tilde{\psi}|V)g(V)] = 0.$$

This concludes that $\psi(P_\lambda)$ is differentiable at P_λ relative to $\dot{\mathcal{P}}_{P_\lambda}$ and $\tilde{\psi}$ is the efficient score; that is, (1) and (2) hold.

7.4 Proof for Theorem 3.2

Similar to the proof of Lemma 7.6 and 7.7, under Assumption 7(1) or (3), and for any $q_r > 0$,

$$\frac{C + (1 - F)^{\leftarrow}(z)}{(1 - F_\varepsilon)^{\leftarrow}(z^{q_r})} \to 0. \tag{7.8}$$

Next, we consider the integrability of the variance at $+\infty$. With a change of variables,

$$\begin{aligned}\int_0^{+\infty} \frac{1 - F_\varepsilon(\alpha + v)}{f(v)} dv &= \int_0^c \frac{1 - F_\varepsilon(\alpha + (1 - F)^{\leftarrow}(z))}{f((1 - F)^{\leftarrow})(z)^2} dz \\ &\geq \int_0^c \frac{1 - F_\varepsilon(\alpha + (1 - F_\varepsilon)^{\leftarrow}(z^{q_r}))}{f((1 - F)^{\leftarrow})(z)^2} dz \\ &= \int_0^c z^{q_r - 2(\xi_r + 1)} L(z) dz.\end{aligned} \tag{7.9}$$

[7] This is because we can define a functional χ as $\chi(\lambda_t) = \psi(P_{\lambda_t}) = \alpha + th$. Then taking the ordinary derivative of $\chi(\lambda_t)$ w.r.t t, we have

$$\partial_t \chi(\lambda_t) = \langle (1, 0, 0), (h, \eta, g)\rangle = h.$$

Since we can choose q_r to be arbitrarily small, the RHS integral will diverge at 0, which means the variance is ∞.

Under Assumption 7(2), there exists q_r such that $q_r > \frac{\xi_r}{\lambda_r}$ and $q_r - 2(\xi_r + 1) \leq -1$. These two inequalities imply that (7.8) holds and (7.9) diverges to ∞, respectively. This concludes $\mathbb{E}|\frac{Y_i - \mathbb{1}\{V_i > 0\}}{f(V_i)}|^2$ is infinite too. Therefore, for both cases, $\tilde{\psi} \notin \mathcal{L}^2(P_\lambda)$.

If $\tilde{\psi}$ is the unique solution to $\langle A^*\psi, (h, \eta, g)\rangle = h$, then we have shown that $\partial_t \chi(\lambda_t) \notin R(A^*)$, in which $R(A^*)$ denotes A^*'s range and A^* is the adjoint of the score operator A. A^*, A, $\chi(\lambda_t)$, and λ_t are defined in the proof of Corollary 3.1. Then, by Theorem 25.32 of Van der Vaart (1998), we can conclude there is no regular estimator in existence.

What is left to show is that $\tilde{\psi}$ is indeed the unique solution to $\langle A^*\psi, (h, \eta, g)\rangle = h$. We assume that there exists $\hat{\psi}(V, Y)$ which also solves

$$\langle A^*\psi, (h, \eta, g)\rangle = h. \tag{7.10}$$

Then, we aim to show $\pi(V, Y) = 0$, where $\pi(V, Y) = \hat{\psi}(V, Y) - \tilde{\psi}(V, Y)$.

First, (7.10) implies

$$\mathbb{E}\pi(V, Y)\eta(\varepsilon) = 0, \quad \forall \eta \in {}_\phi\dot{\mathcal{P}}_{P_{\lambda_t}},$$

$$\mathbb{E}\pi(V, Y)g(V) = 0, \quad \forall g \in {}_f\dot{\mathcal{P}}_{P_{\lambda_t}},$$

and

$$\mathbb{E}\dot{l}_\alpha(V, Y)\pi(V, Y) = 0.$$

Thus, there exists some constants C_1 and C_2, such that

$$\mathbb{E}(\pi(V, Y)|\varepsilon) = C_1 + C_2\varepsilon, \ \mathbb{E}(\pi(V, Y)|V) = 0, \text{ and } \mathbb{E}\dot{l}_\alpha(V, Y)\pi(V, Y) = 0.$$

Further note

$$\mathbb{E}(\pi(V, Y)|\varepsilon) = \int_{\varepsilon-\alpha}^{\infty} \pi(v, 1) f_V(v) dv + \int_{-\infty}^{\varepsilon-\alpha} \pi(v, 0) f_V(v) dv,$$

in which $f_V(\cdot)$ is the density of V. Taking derivatives on both sides w.r.t. ε and letting ε range over $\Re$, we have $\pi(t, 0) - \pi(t, 1) = \frac{C_2}{f_V(t)}$ for any $t \in \Re$. Then,

$$0 = \mathbb{E}\dot{l}_\alpha(V, Y)\pi(V, Y) = \mathbb{E}\phi(V + \alpha)(\pi(V, 1) - \pi(V, 0)) = C_2\mathbb{E}\frac{\phi(V + \alpha)}{f_V(V)} = C_2.$$

This implies $\pi(t, 1) = \pi(t, 0) = \pi(t)$. At last,

$$0 = \mathbb{E}(\pi(V, Y)|V) = \pi(V),$$

i.e., $\pi(V, Y) = 0$. This concludes the uniqueness of $\tilde{\psi}$, and thus, the whole proof.